Group2_1 - Lab Exercises 1 and 2

Athika Fatima - 101502209
Dev Chetal - 101459557
Sreejita Chowdhury - 101590107
Dhruv Jayal - 101503569
Panchasara Mohitkumar Jayeshbhai - 101567404
Shreya Lingwal - 101583877
Tri Thanh Alan Inder Kumar - 101413004
Chirag Vaghasiya - 101505362
Vrunal Patel - 101574257

1. Business Analysis: Context and Framework

a. Read and understand the spreadsheet's data

The dataset from **Sample - Superstore** contains 9994 rows and 21 columns, representing sales transaction data including order details, customer details, product categories, regional information, and performance metrics such as sales, discounts, and profits. Key columns include:

- **Order Information:** Order ID, Order Date, Ship Date, Ship Mode
- **Customer Details:** Customer ID, Customer Name, Segment
- **Location Data:** Country/Region, City, State, Postal Code, Region
- **Product Information:** Product ID, Category, Sub-Category, Product Name
- **Sales Metrics:** Sales, Quantity, Discount, Profit

---

b. Indicate the sector/industry and key entities impacting the reporting process

- **Sector:** Retail and Consumer Goods
- **Key Entities:**
    - **Customers:** Essential for understanding demand and preferences.
    - **Products:** Impact revenue and profitability across categories and sub-categories.
    - **Orders:** Track transactions, sales performance, and applied discounts.
    - **Regions and Cities:** Enable performance insights and geographic analysis.
    - **Sales Metrics:** Key indicators for generating performance KPIs.

---

c. Define the target audience for the reports

- **Operational Report:** Regional managers, sales teams, and operational analysts focused on day-to-day performance.
- **Executive Report:** Senior executives and decision-makers responsible for strategic business planning.

---

d. Indicate the intended use of each report

- **Operational Report:**
  - Monitor and control daily and weekly sales performance.
  - Identify operational issues such as missed quotas, negative profit margins, and regional underperformance.
- **Executive Report:**
  - Support decision-making through high-level performance analysis.
  - Compare revenue and profitability across regions and product lines.
  - Evaluate growth opportunities and strategic alignment.

---

e. Indicate publishing, distribution frequency, maintenance, and use of the reports

- **Operational Report:**
  - **Publishing Frequency:** Weekly or monthly.
  - **Distribution:** Regional managers and sales teams.
  - **Maintenance:** Periodic updates using clean, validated data.
- **Executive Report:**
  - **Publishing Frequency:** Monthly or quarterly.
  - **Distribution:** Senior management.
  - **Maintenance:** Updated periodically with key metrics adjusted as per organizational goals.

---

f. Indicate any assumptions you need to produce correct deliverables in a timely manner

- The data in the spreadsheet is complete and up to date.
- KPIs such as sales growth and profit margin can be derived from existing fields.
- Stakeholders have communicated their reporting needs.
- Any missing values in critical columns (like sales or dates) can be reasonably imputed.

---

g. Indicate the information that both Operational and Executive Reports will include

- **Operational Report:**
  - Daily/weekly sales performance by region, city, and salesperson.
  - Product-wise breakdown including category and sub-category performance.
  - KPIs such as total sales, profit margins, quota attainment, and sales growth.
- **Executive Report:**
  - Aggregated monthly/quarterly performance data.
  - Key KPIs: total revenue, profit, sales growth, and top-performing products/regions.
  - Comparison of current performance with previous periods.

---

2. Technical Analysis

h. Identify and document completeness, inconsistencies, redundancies, and duplicates

- **Completeness:** Checked for missing values in key columns such as Sales, Order Date, Region.
- **Inconsistencies:** Verified consistency in date formats, and categorical values like Region and Category.
- **Redundancies:** Checked for duplicate rows based on Order ID and Product ID.
- **Duplicates:** Duplicates were identified and removed using Python.

Python Code:

```
# Check for missing values
data.isnull().sum()

# Drop duplicate rows
data = data.drop_duplicates()
```

---

i. Indicate any additional calculations (formulas/algorithms), KPIs, new entities, and attributes

- **KPIs Calculated:**
  - **Profit Margin (%)** = (Profit / Sales) × 100
  - **Sales Growth (%)** = ((Current Sales - Previous Sales) / Previous Sales) × 100
  - **Discount Impact** = Discount × Sales

- **Return Rate (%)** = (Number of Returned Orders / Total Orders) * 100
- **Average Order Value (AOV)** = Total Sales / Number of Orders

Python Code:

```
# Profit Margin (%)
data['Profit Margin (%)'] = (data['Profit'] / data['Sales']) * 100

# Sales Growth (%)
# Assuming 'Order Date' is used to calculate sales growth over time
data['Year'] = data['Order Date'].dt.year
data['Month'] = data['Order Date'].dt.month
monthly_sales = data.groupby(['Year', 'Month'])['Sales'].sum().reset_index()
monthly_sales['Sales Growth (%)'] = monthly_sales['Sales'].pct_change() * 100
print("\nMonthly Sales Growth Calculated.")

# Discount Impact
data['Discount Impact'] = data['Discount'] * data['Sales']

# Return Rate (%)
# Merge with the Returns sheet to identify returned orders
returns = pd.read_excel('/content/sample_data/Sample - Superstore - Copy.xls',
sheet_name='Returns')
data['Returned'] = data['Order ID'].isin(returns['Order ID']).astype(int)
total_orders = data['Order ID'].nunique()
returned_orders = data['Returned'].sum()
return_rate = (returned_orders / total_orders) * 100
print(f"\nReturn Rate: {return_rate:.2f}%")

# Average Order Value (AOV)
total_sales = data['Sales'].sum()
total_orders = data['Order ID'].nunique()
aov = total_sales / total_orders
print(f"Average Order Value (AOV): ${aov:.2f}")
```

---

j. Clean/transform the data source for further use and data analysis

- **Duplicates Removed:** Dropped duplicate rows.
- **Missing Postal Codes:** Filled missing postal codes with -1 (placeholder).

- **Standardized Formats:** Cleaned categorical columns (e.g., region, product names) and standardized dates.

Python Code:

```
# Fill missing postal codes
data['Postal Code'] = data['Postal Code'].fillna(-1)

# Standardize date formats
data['Order Date'] = pd.to_datetime(data['Order Date'])
```

---

3. Report Form Designs

Operational Report Form

- **Title:** Operational Report
- **Columns:**
  - Date (Daily/Weekly)
  - Region
  - City
  - Product Category
  - Sub-Category
  - Total Sales
  - Quantity Sold
  - Discounts Applied
  - Profit Margin (%)
  - Sales Growth (%)

Example Layout:

Date	Region	City	Product Category	Sub-Category	Total Sales	Quantity	Discount	Profit Margin (%)	Sales Growth (%)
2018-01-03 00:00:00	Central	Houston	Office Supplies	Paper	16.448	2	0.2	33.75	0

2018-01-04 00:00:00	Central	Naperville	Office Supplies	Storage	272.736	3	0.2	-23.75	1558.171
2018-01-04 00:00:00	Central	Naperville	Office Supplies	Binders	3.54	2	0.8	-155	-98.702
2018-01-04 00:00:00	Central	Naperville	Office Supplies	Labels	11.784	3	0.2	36.25	232.8814
2018-01-05 00:00:00	East	Philadelphia	Office Supplies	Art	19.536	3	0.2	25	65.78411
2018-01-06 00:00:00	South	Henderson	Office Supplies	Paper	6.54	1	0	46	-66.5233
2018-01-06 00:00:00	South	Athens	Office Supplies	Art	12.78	3	0	41	95.41284
2018-01-06 00:00:00	West	Los Angeles	Office Supplies	Paper	19.44	3	0	48	52.11268
2018-01-06 00:00:00	South	Henderson	Technology	Phones	755.96	4	0	27	3788.683
2018-01-06 00:00:00	South	Henderson	Office Supplies	Art	5.48	2	0	27	-99.2751

Executive Report Form

- **Title:** Executive Report
- **Columns:**
    - Region

- Period
- Total Sales
- Total Profit
- Profit Margin (%)
- Sales Growth (%)
- Top-Performing Products (based on profit or sales volume)
- Discount Impact
- Return Rate
- Average Order Value

Example Layout:

Region	Year-Month	Total_Sales	Total_Profit	Profit_Margin_Percent	Sales_Growth_Percent	Top-Performing Products	Discount Impact	Return Rate (%)	Average Order Value (AOV)
Central	2018-01	1539.906	118.4902	3.275132	369.8877	Storage	411326.5	25.97424	233.6294
Central	2018-02	1244.538	290.2842	-29.0833	95.19962	Accessories	411326.5	25.97424	233.6294
Central	2018-03	5827.602	-274.047	1.334524	2202.795	Tables	411326.5	25.97424	233.6294
Central	2018-04	3712.34	229.3787	20.70578	661.3353	Tables	411326.5	25.97424	233.6294
Central	2018-05	4048.506	-506.017	-18.1342	231.6645	Chairs	411326.5	25.97424	233.6294
Central	2018-06	9646.299	1246.122	0.959456	736.7985	Chairs	411326.5	25.97424	233.6294
Central	2018-07	20610.49	-22964.3	-79.409	781.7158	Binders	411326.5	25.97424	233.6294

Central	2018-08	3022.183	628.4268	-2.53968	1178.723	Storage	411326.5	25.97424	233.6294
Central	2018-09	34408.69	1422.797	-19.7983	3688.467	Machines	411326.5	25.97424	233.6294
Central	2018-10	8965.757	1341.241	-29.317	4709.312	Phones	411326.5	25.97424	233.6294